档案数据化过程中语义组织的内涵、特点与原理解析

■ 祁天娇 冯惠玲
中国人民大学信息资源管理学院 北京 100872

摘 要：[目的/意义]档案数据化阶段，档案利用与服务需要满足用户在数据层级的需求，突破页面级阅读和文件级利用的限制，这就要求在组织环节构建起语义层级的档案组织新模式，以实现对档案内容、背景与结构数据的细颗粒分析与挖掘，面向档案资源增值、开发与智能化知识服务做好资源、方法与技术的准备。[方法/过程]采用文献调研与案例分析等方法，立足档案数据化阶段特征，分析档案语义、语义关联和语义组织的基本内涵，比较分析档案与其他信息资源在语义组织过程中的区别与特性，探索在语义完整、链式关联以及网络多维原则下开展档案语义向内组织与向外组织的基本原理。[结果/结论]档案语义组织是基于数据的含义与关联开展的档案组织新模式，旨在从档案资源的内容、背景与结构数据中发现语义与语义关联。档案语义组织是实现档案数据化的核心环节，是实现档案机器可理解、机器可操作的关键一步。通过档案语义组织，原本离散、分布、领域依赖的档案内容、背景与结构数据能够含义明确化、编码形式化、关系链接化，档案数据得以被机器可理解、可操作，档案自动化关联组织、存储与提供利用成为可能，从而最终支持基于人机交互、机机交互的档案资源智能化获取、利用与服务。

关键词：档案 档案数据化 语义 语义关联 语义组织
分类号：G271
DOI：10.13266/j.issn.0252-3116.2021.09.001

1 引言

经过20年的存量档案数字化与增量电子文件归档管理，数字态档案资源储量已非常丰富[1]，但现有的档案组织手段仅能支持基于页面阅读的文件级档案利用与服务，无法实现档案内容的可理解以及内容、背景、结构的关联性利用，业务驱动下的档案如何与文件生成背景建立联系，为业务端提供信息“反哺”以支持业务决策，也是档案管理向前突破的关键问题。原国家档案局局长杨冬权曾指出：“我想利用档案，不用我去找，自动地就能推送过来，这就需要做一件更重要、工作量也更大，意义和价值更大的事情，那就是把档案数据化”[2]。这指出当前面临语义网、人工智能等新计算机技术冲击时，档案数据化所引导的新转型之路将要走向的目标：超越文件级查找和页面阅读的档案利用局限，为用户提供更加智能化的档案服务。档案数据化是档案数字化的更高阶段。档案数字化将档案信息由模拟或物理信号转变为“0”“1”的数字形式，通过扫描、计算机文字处理等将档案文本中的固化对象转化成离散的比特（bits），存储在计算机系统或数据库中而非纸质媒介中[3]。档案数据化则将“0”“1”等离散的比特（bits）进行再组织，形成结构化的、标准化的、开放性的、可通用的数据对象，并基于数据对象的不同形态与类别开展相应的机器操作活动[4]。由此可见，档案数据化的关键在于将零散的比特（bits）组织成有含义的、有关联的数据集合，即数据的组织。这种数据组织的核心目标在于使机器可理解数据的含义，并基于此实现机器对数据的自动化操作。因此从数据的含义层面开展数据组织就尤为重要，这就是本文所探讨的语义组织。

信息资源组织过程中的语义技术应用，是近年来图档博领域的一个研究热点。现有研究主要反思了信息组织的不足，主张向更细颗粒的信息组织、更智能化的信息传播与应用转移，强调信息资源的分类和描述等要向语义揭示及关系发现的深层次发展。因此，“语义组织”也成为一个被信息资源管理领域高频讨论的学术概念。但遗憾的是，当前“语义组织”这一概念尚未有统一表述，有学者从语义组织的对象入手，认为语

作者简介：祁天娇（ORCID：0000-0001-5595-4774），博士，博士后，E-mail：qtjjoy@163.com；冯惠玲（ORCID：0000-0003-4800-1259），教授，博士。
收稿日期：2020-11-13 修回日期：2021-02-02 本文起止页码：3-15 本文责任编辑：杜杏叶

义组织包括“语义描述、本体转化、发布为关联数据”三个层面的基本内容[5]；也有学者针对不同领域的不同信息资源类型开展具体的语义组织探索，包括：面向文化传播与传承的非遗多媒体资源的本体模型构建和分层语义描述[6]、面向政府决策的舆情信息语义组织[7]、面向 e-Science 的科学数据语义组织研究[8]以及科技报告语义关联研究[9]等。在语义技术应用路径层面，现有研究成果的基本思路是对信息资源进行知识抽取、本体构建、知识图谱和本体检索等[10]。但因为支持语义组织的技术标准和方法如语义网、知识工程、人工智能技术等处在快速更新之中，学界关于信息资源语义组织的研究处在快速的概念扩展和技术融合阶段，也同时出现了术语或概念混淆使用等问题。例如，元数据、词表、本体、关联数据等概念反复出现在众多研究成果中，但各中原理及其与信息资源组织变革的关系尚未有清晰解释。

在信息资源大领域的带动下，档案领域也开始出现语义技术应用方案。例如美国 FamilySearch. org 和 ancestry. com 等家谱网站利用本体技术重构家谱档案数据，利用时空关系等多维语义关系的建立来揭示隐藏在家谱档案数据中的人物关系和其他知识，并为用户提供多维检索。日本神奈川大学非书写遗产中心将部分民俗用具资料数据化，并基于本体技术构建民俗用具数据库[11]。2011 年，法国国家档案馆以 RDF 格式发布了叙词表，为用户提供关联数据的语义查询服务[12]。我国学者也提出应用相关技术来改进档案组织，例如裘丽认为应利用语义网技术实现数据转换、描述、分类，利用智能 Agent 技术为模糊性用户进行服务信息的整合优化[13]；马寅源基于 SWOT 分析法，分析了关联数据方法在档案知识服务中应用的影响因素和对策，认为关联数据的应用是未来档案知识服务的重要方向[12]。但“档案语义组织”这一概念的准确含义以及语义组织在档案领域的独特内涵，学者们虽有讨论但尚未有定论。例如有学者从“组织什么”的问题出发，认为档案的语义组织主要包括档案信息资源内容体系和知识体系的语义关系组织、词汇体系与元数据体系的映射关系组织等方面[14]。有的学者则从“如何组织”的问题出发，认为档案的语义组织流程包括元数据语义转换、档案数据语义分析与表述、语义组织与存储、语义检索与服务等方面[15]。这样的内涵实际上是操作路径，档案的语义、语义关联和语义组织到底是指什么，尚无明确答案。

在这样的背景下，本文以档案语义组织是什么、如何实现为基本研究问题，采用文献调研与案例分析等方法，剖析档案组织发展过程中的语义传统，并将档案语义组织与档案实体组织、档案信息组织相比较，在数据化背景下定义档案语义组织的内涵，并基于档案资源的特性探讨语义技术应用于档案组织中的原理与原则，为档案数据化过程中数据组织的核心问题提供语义层面的解决方案。

2 档案组织中的语义组织传统

档案语义组织是语义技术在档案组织中的新应用，是数据化过程中出现的档案组织新思维与新方法。但语义组织并不是凭空出现的，档案组织长期的发展过程中，也形成了潜在的语义组织传统，能够为新时期的档案语义组织发展奠定良好基础。

2.1 档案实体组织中的语义组织传统

档案实体组织是针对档案物理实体（载体）的组织，其目的是实现馆藏档案实体的序化。在我国，档案实体组织借鉴了前苏联“国家档案全总条例”的约定[16]，以全宗原则为核心，再依据档案实体特征分类立卷。其基本环节包括划分全宗、全宗内档案分类、立卷以及卷内排列，从而对档案的来源、时间、内容和形式特征等进行分类。和宝荣、陈兆祦、松世勤提出的档案整理工作的内涵影响至今，大体分为“系统化和基本编目两大部分”，并以“按照文件之间的历史联系整理档案”为原则[17]。这里的“历史联系”是指档案案卷内文件之间包括来源方面、时间方面、内容方面和形式方面的联系，档案实体大多基于档案来源（组织机构）、时间（年度）、内容（事由）和形式（种类）分类，以案卷形式被整理排架，其中内容（事由）的分类就是在档案内容主题的语义层面的分类组织方法。

2.2 档案信息组织中的语义组织传统

档案信息组织是针对档案检索信息的组织，其目的是实现档案检索信息的序化。档案信息组织经历了较长历史，分类卡片、比孔卡、穿孔卡、边缘穿孔卡等都是手工管理环境下重要的档案信息组织工具，档案信息组织的结果一般表现为档案目录、索引、编研成果等二、三次文献。分类法和主题法是档案信息组织最重要的两大方法，而这两种方法都是在语义层面、从档案内容分析入手进行档案标引和检索的组织方法，无论是分类法中的类号和类名，还是主题法中的主题词，本质上都是对档案所反映概念的表达。其中，档案信息分类体系是一种列举已知类目并逐级展开的层累制的号码检索体系，以概念的划分和概括的原理为基础，反

映档案内容的从属派生与平行关系[18]。为了在馆藏系统化的基础上统一档案检索的分类方法，突破档案实体分类中的年度、组织机构、类型等形式特征分类限制，《中国档案分类法》(简称"中档法")提出了"以统一分类原则与标记制度为前提，以职能分工为分类标准和依据，结合体系分类法和分面组配法并具有半分面组配性质"[19]的档案分类体系。这种"职能分工"的分类原则实际上是从档案所参与的社会职能的角度来分析档案内容的语义。另一方面，档案信息组织的主题法通过自然语词来描述档案中的各种概念，并将各种概念按字顺排列。与分类法的层级组织方式不同，主题法采用分面组配的方式来揭示档案主题，并以规范化的自然语词作为标引和排检依据，实质上是一种档案主题词典[18]。这种主题词典在当下具有很强的转变为"数据词典"——本体的潜力。

2.3 网络档案信息组织中的语义组织传统

无论是档案实体组织中的内容分类，还是档案信息组织中的分类法与主题法，虽然都是对档案内容、主题的语义层面的组织，但其类目或主题词都是以人工标记或自然语言表达的，是人可理解的。这种语义组织方法主要面向人而非机器。以机器可理解的方式开展档案组织首先萌发在网络档案资源的管理中。网络档案资源组织需要对互联网上大量分散无序的档案信息进行筛选、排序、著录、标引、分析、存储、利用，使之形成系统化的结构[20]。传统依赖手工和专家的组织方式无法应对海量网络档案资源的处理需求，自动化组织手段包括自动分类、自动标引、自动编制和管理分类表、词表，自动编制目录、索引、文摘以及自动搜索网上信息源等，能够更有效地处理文本、图形、图像、声音、动画、视频等复杂多媒体信息，而超文本链接能够将这些复杂资源关联起来，形成更大范围的资源网络[21]。但是这种网络档案资源的网状组织和关联，仍然是文件层级的，超文本链接本身并不具有语义，网络档案资源间为什么具有这样的关系和链接，仍然需要资源利用者自行判断。

借助网络档案资源组织的探索，越来越多的学者开始关注语义网的发展，并借助语义技术改进以元数据为核心的档案资源组织方法。其中，学者们讨论最多的就是如何利用本体、关联数据等语义网技术来进行档案信息的标识、描述和推理[22]，解决档案信息与档案信息系统的异构问题[23]，或者将其应用到数字档案馆中，以解决数字档案资源的知识关联[24]、语义互操作[25]、跨媒体语义检索[26]和语义聚合[27]等问题。从研究结论上看，现有研究成果已经提出了档案组织"应该用本体、关联数据"，为本文的研究提供了重要基础，但现有成果更多是在信息资源的普遍框架下讨论语义技术的应用，少有成果从档案资源的特殊属性出发探讨档案语义组织区别于一般信息资源语义组织的内涵，并独立探讨档案领域的语义组织方案。

由此可以看出，档案语义组织在档案实体组织、档案信息组织以及互联网环境中都有特定的内涵，在档案数据化的整体趋势下，档案语义组织将更明确地、从含义层面针对档案本身的数据和描述档案的数据开展组织，将不再囿于分类法或主题词表的制定，而更灵活地包容多种语义技术，以形式化语言面向机器理解实现语义层面的组织。

3 档案语义组织的内涵

特里·库克曾说："通过向世界展示如何避免淹沉在无意义的数据'海洋'里和如何探求相互关联的意义或知识，来重新肯定我们专业的适用性。"[28]这种寻找关联的思维，深刻影响了过去20年来档案领域的理论研究与实践探索，如今语义技术与数据管理思维的新潮，将"事务"与"关系"重新拉回数据组织的核心位置[29]。在新的计算机环境中讨论"相互关联的意义或知识"，就是讨论相互关联的具有明确含义的数据以及经过关联组织后形成的知识网络，这是本文探讨档案语义组织的本质追求。因此，档案语义组织首先要从数据的含义即语义、具有明确含义的数据的关联即语义关联，以及如何组织数据的含义与关联即语义组织的内涵谈起。

3.1 档案语义的内涵

语义是指数据的含义，需要遵循一定语法的形式化语言来表达，即这种含义可以被机器所理解，一般基于自然语言描述的数据的含义则需要转化为形式化语言的表达。档案的语义是指所有档案本身的数据和描述档案的数据的含义，包括档案内容数据、背景数据和结构数据的含义，与传统档案文本内容的含义或元数据的含义不同，档案的语义使用形式化语言表达，含义明确且机器可理解。

3.1.1 档案内容的语义

档案的内容是指档案中所包含的表达作者意图的信息[30]。档案内容一般采用自然语言表达，其含义取决于文件形成时作者使用的词语和句法结构，依赖一定的语言体系和上下文关系，可能因语种和语境的变化而变化。因此，识别档案内容的语义，就是要识别文

本中的词语及其指代的概念之间的对应关系。为了获取内容语义，一般会采用语义标注（标引）的方法，借助本体等工具识别文本中的概念，使以文件为单元的信息组织发展为以概念为单元的语义组织[31]。

档案内容的语义集中体现在时间、人物（机构）、地点、事件（主题）和实物五大方面，可以通过标注文本中这五大要素加消歧的方法，来获取档案内容中最重要的语义。如在台湾历史数位图书馆中，有标题为《立杜卖尽根璞园字》的一份档案[32]其文本中的地名如"拣东上堡七份庄"、人名如"吴阿旺"等语义被标注出来，在经过形式化编码表达后，这些内容语义就可以与其他档案中的相关语义进行关联组织，抽取出这些档案中关于同一时间、人物、地点和事件的知识。当档案资源文本内容语义的标注颗粒越细时，内容中语义关系的揭示程度就会越高，档案资源文本内容中所蕴含的知识被发现、聚合、挖掘的深度与效果就会越好[33]。

3.1.2　档案背景的语义

档案的背景是指档案所处的环境[30]。任何只有内容信息而不具备背景信息的档案都是不完整的、缺乏凭证性的。任何档案都有其机构背景、业务背景、程序背景和文件背景[34]。档案背景的语义就是指描述档案机构背景、业务背景、程序背景和文件背景的数据的含义。其中，机构背景是指档案生成者所属的机构体系；业务背景是指生成档案的业务职能、活动和事务；程序背景是指文件生成、转递、归档与管理的程序；文件背景是文件所属的档案全宗或档案汇集内与其他文件之间的关系。

背景驱动是档案资源区别其他信息资源的一个重要特征。档案资源不是"静态"资源，在其生命周期演变过程中，背景信息深刻影响着档案资源的内容与结构。档案背景语义的识别和获取，对于建立起档案文本与档案来源机构、业务、程序和文件汇集之间的相关关系，拓展档案资源文件层、汇集层甚至全宗层的外部关联，具有极为重要的作用。例如图 1 所示，"台湾历史数位图书馆（THDL）"根据档案的文件背景语义，识别历史公文在"上传下达"中的环节与作用，提供每份历史档案与其他档案的相关关系"另类视窗"[35]，为档案检索用户提供更多历史档案的浏览推荐。

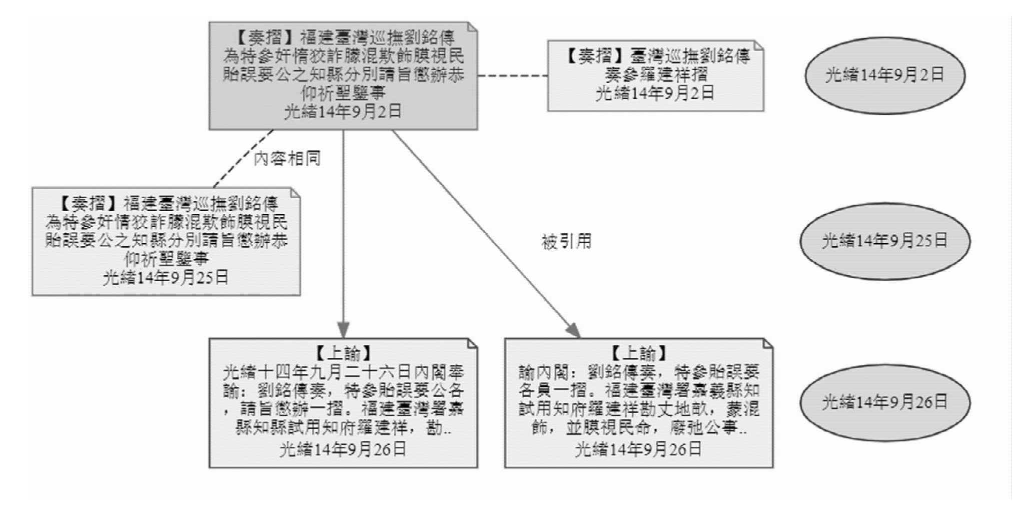


图 1　THDL 通过识别文件背景语义提供"另类视窗检视"服务

3.1.3　档案结构的语义

档案的结构是指档案内容信息的组织方式与表达方式，其中组织方式包括正文和附件，表达方式包括格式、载体、版本等[30]。档案结构的语义就是描述档案结构的数据的含义。在传统纸质档案资源中，档案内容的语义与档案结构的语义是不可分离的，但随着档案数字化以及电子文件的发展，档案的内容与结构逐渐分离，且相互之间的影响度逐渐缩小。例如在很多信息管理系统中，某类文件内容数据可填写、结构数据则模板化自动生成，最终呈现出认可阅读且具有固化结构的文件，也就是说文件的页面布局等结构语义已经由机器自动设定且机器可理解。在这种背景下，档案结构的语义更容易被独立识别并获取，一般体现在描述档案资源长期保存信息的数据的含义，包括档案的格式、版本、载体数据等。在数据化状态中，结构语义对于定义文件所处生命周期阶段必不可少，不同版本和格式的文件可能处于不同的机构、业务、程序和文件背景中，也可能包含不同的文本内容信息。因此，档

案结构的语义往往是建立起档案资源内容语义与背景语义之间关系的重要桥梁。

通过档案内容、背景和结构语义之间的关联,档案之间新的关联也将建立起来。如图2所示:

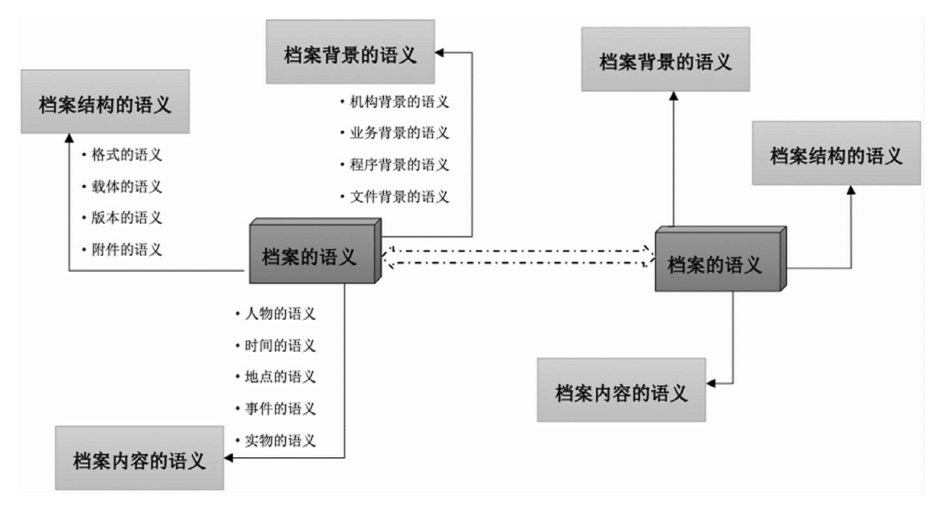


图2　档案语义的内涵及其关系

3.2　档案语义关联的内涵

档案的语义蕴藏在档案的内容、背景与结构数据中,而这些数据不仅存在于文件中,也存在于档案汇集的各个层级中,且不同层级的内容、背景与结构数据在含义和性质上可能具有继承或其他关系,这些不同层级的内容、背景、结构语义之间的关联,就能够建立起不同层级档案之间的关联,从而形成多层级的数据和语义网络,如图3所示:

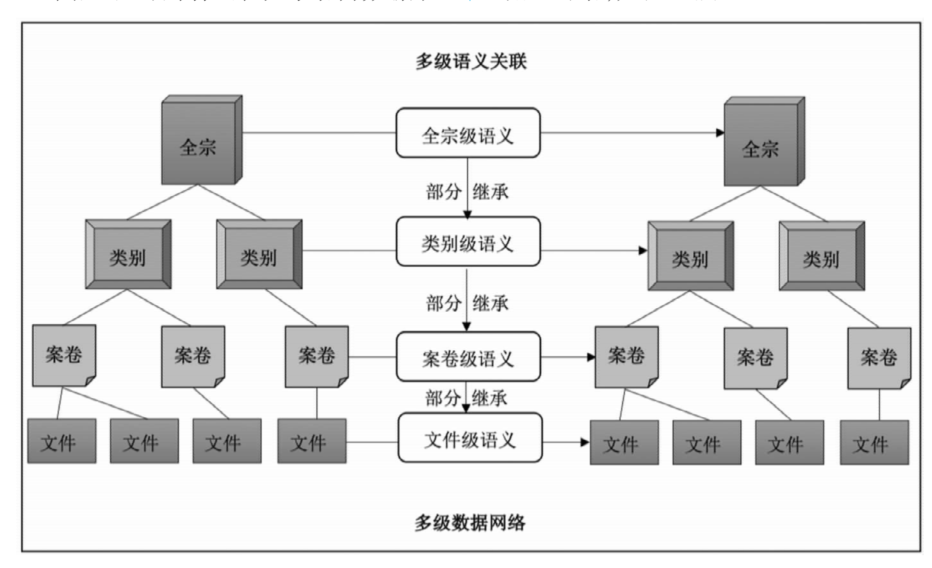


图3　档案资源的多级语义关联

3.2.1　多级著录引发的多级关联

档案的多级著录思想来源于 Oliver W. Holmes 的五级档案整理理论,即分别在档案仓储(depository)、文件汇集(recordsgroup)、系列(series)、案卷(file unit)、文件(document)五级开展档案资源的著录和整理工作[36]。1992年第十二届国际档案大会确立了现代档案著录的"马德里原则",即来源原则、尊重全宗原则和反映管理级次原则,其中反映管理级次原则就要求档案著录必须充分反映"全宗-分全宗-类别-案卷-文件"这样的等级层次[37]。ISAD(G)明确提出了要开展全宗(子全宗)-系列(子系列)-文件-实体四个层面的多级著录[38],从而为纵向的全宗内编目和检索、横向的层级间相关性检索提供支持。多级著录能够提供多层、多维、完整的档案描述,为档案不同层级之间的关联建立奠定了基础[39]。多级著录也为档案检索提供了多级入口,任何一个层级的著录数据都可以向上级追溯或向下级延伸,以获取更高或更低级别的档案著录数据,同时也可以向左右扩展获取相关档

案的关联检索。

我国《档案著录规则》(DA/T18－1999)仅规定了文件级,案卷级的档案内容、背景与结构著录项目与格式,对全宗和类别级的著录则几乎没有标准的数据定义、项目参考和格式限定。在这样层级单薄的著录体系下,档案的语义大多来源于文件层级的内容、背景与结构数据,而无法从案卷级、类别级、全宗级获取更多的语义,也就无法在四个层级以及层级之间都建立起丰富的语义关联。

因此,要获得更完整的档案资源语义关联,就要完成多个层级的档案资源语义著录工作。这包括获取多级著录数据和分析多级著录数据的语义两个层面:

(1)对全宗－类别－案卷－文件四个层级的档案资源开展多级著录,设定每个层级的著录项目与格式,明确高层级著录项目与低层级著录项目之间的继承关系、部分继承关系,避免多层级著录的重复,同时注重不同层级所含内容、所属背景、所具结构的特殊性及其专门著录。

(2)对全宗－类别－案卷－文件四个层级的多级著录数据进行语义描述,分析不同层级著录数据在内容、背景、结构上的语义聚类与关联关系。

3.2.2 多级关联形成的多级网络

经过多级著录后,全宗－类别－案卷－文件四个层级的档案语义就能够被进一步分析、抽取和关联起来,从而建立起纵横两个方向的语义关联网络:

(1)纵向网络是指档案的全宗－类别－案卷－文件四个层级之间语义关联的建立,这种关联是基于上下层级档案内容、背景、结构语义的部分继承关系所建立的。一般来讲,纵向网络是档案语义网络的主线,是对档案编目传统的继承,但改变了严谨的根系树状结构,而能够提供上下级、跳级、单级等多种档案描述与检索的扩展与缩减。

(2)横向网络是指档案的全宗－全宗、类别－类别、案卷－案卷、文件－文件四个层级的同级语义关联的建立,这种关联是基于同级档案在内容、背景与结构语义上的相关关系。一般来讲,这种相关关系需要外部开放资源作为关联桥梁,例如两个不同档案汇集通过"机构名录"这一外部资源,建立起两个档案汇集所属机构在职能上的上下游关系,因而进一步在机构全宗的背景语义上建立关联,从而实现从一个机构的全宗到另一个机构的全宗的语义链接。

基于纵横两向基本语义关联网络,更多的交叉层级的档案语义网络也可以建立起来,从而能够提供更多的档案检索点和服务入口。

3.3 档案资源语义组织的内涵

档案的语义组织是指将档案内容、背景与结构数据含义明确化、编码形式化、关联链接化的过程,包括识别、理解、分析和表达档案的语义,并建立起多级档案语义之间的关联这两大部分。任何语义也不可能脱离其他语义而独立发挥价值,根据语义间的关系对其进行分类、聚类、关联等,就可以形成描述客观世界中的概念或知识。因此,语义组织的本质就是分析语义之间的关系、建立语义之间的关联。根据档案语义来源于档案汇集内或档案汇集外,档案语义组织可以分为向内语义组织和向外语义组织。

3.3.1 向内语义组织

向内语义组织的"内"是指一个特定的档案汇集之内。被组织的语义来自同一个数据源中的数据,不涉及跨档案汇集或跨数据源的关联或集成问题。向内组织也是建立档案纵向语义关系网络的过程。对于大多经数字化的历史档案来说,为历史研究或公共记忆构建服务是主要目标,其档案向内语义组织主要是指某一历史研究主题或公共记忆方向的档案汇集内档案文本内容的语义组织,即根据文本内容中人物(机构)、时间、地点、事件或者实物语义之间的关联,建立起某一主题下档案汇集内不同文件之间的关联,形成关于这一主题的完整、详细的内容网络。对于大部分原生电子档案来说,提供业务凭证、支持业务决策为主要服务目标,其档案向内语义组织主要是指某一机构档案汇集(或机构全宗)内"全宗－类－文件－组件"的多层次的内容语义、背景语义和结构语义的关联组织。如图4所示,每一层的语义组织都会涉及到内容语义、背景语义和结构语义之间的关联,最终构建起四层网状结构。

3.3.2 向外语义组织

向外语义组织的"外"是指一个特定的档案汇集之外,即超越档案汇集的限制向更多外部数据源寻找并关联相关语义,包括不同档案汇集之间的语义关联、档案汇集与其他类型数据汇集之间的语义关联两种基本类型。向外语义组织将档案汇集视为更多领域档案汇集中的一个组织部分,强调基于领域知识对不同档案汇集甚至不同数据汇集进行集成和关联组织,从而形成对更广泛的领域知识的描述、开发和利用。档案语义向外组织也是建立档案横向语义关联网络的过程。档案语义向外组织的关键是通过机器可理解的链

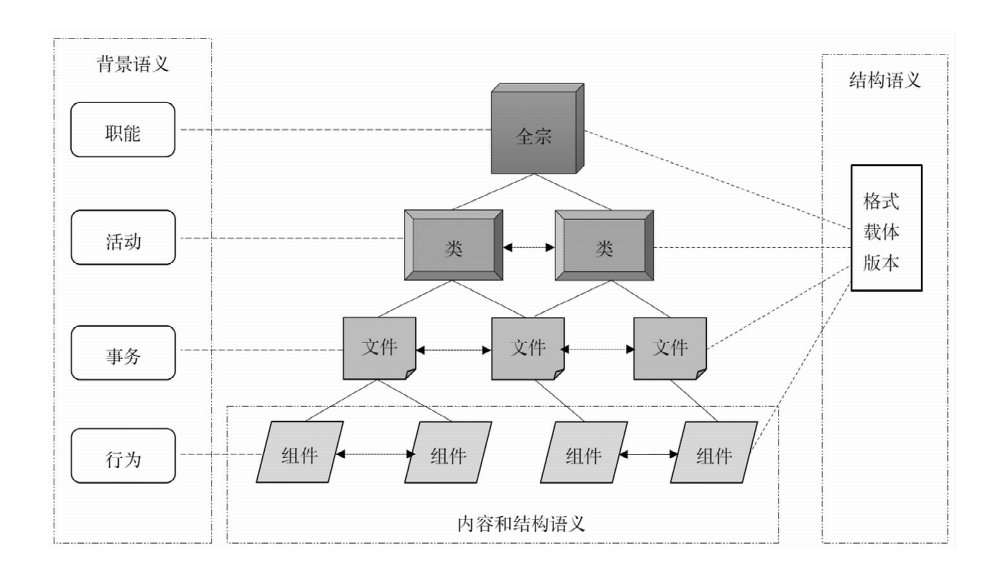


图 4　面向业务的档案向内语义组织

接，建立起不同数据源数据间的关联，实质是通过语义集成实现语义关联。对于大部分业务驱动的电子档案来说，向外语义组织意味着超越机构职能和档案来源限制的，相关机构、职能、档案汇集之间的社会关系网络的综合建立。正如图 5 所示，原本来源于不同机构的档案汇集相互独立，因为机构、职能和业务之间的上下游关系，并在产生关系的同时产生相应的业务凭证，而最终形成因业务领域相关而聚集的跨越机构、职能的系列档案，使得原本分散独立的档案汇集之间具有了关系。这就要求在组织档案时，不能仅向内组织机构职能内部的业务活动、事务及其生成的各类档案，还要向外组织相关的机构名录、职能列表、业务活动记录、事务日志、行为数据等，通过与公开的外部资源的关联，建立起不同档案汇集在背景语义尤其是机构语义、业务语义方面的关联性，建立起机构档案汇集内的档案与汇集外的档案的关系，形成对整个业务领域内职能、活动的完整梳理，最终为整个业务领域而非仅为某个职能或活动提供档案资源服务。

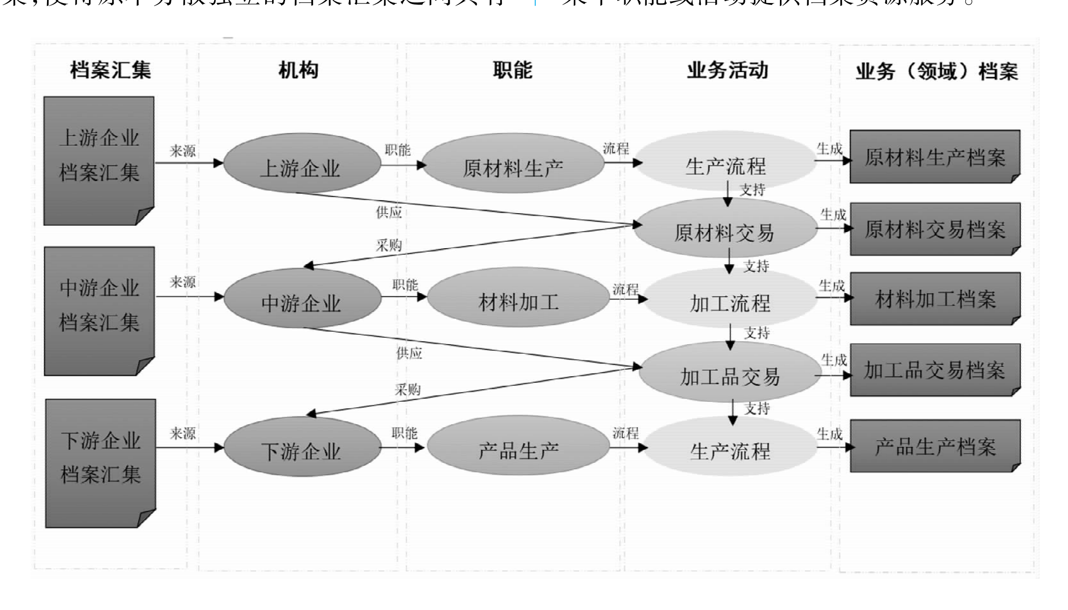


图 5　业务驱动的档案资源向外语义组织

4　档案语义组织的原理

档案语义组织是借鉴语义网中信息资源组织的基本原理，对档案的内容语义、背景语义、结构语义进行向内组织和向外组织的过程，是推进档案数据化的核心环节。档案语义组织的原理围绕三大核心问题的解

决:档案语义从何处来?档案语义如何关联?档案语义和语义关联如何为机器所理解?

4.1　结构化的语义来源

语义是数据的含义,语义来源于数据,而机器可理解和可操作的语义主要来源于那些被数据模型严格定义的结构化数据[40]。对于以非结构化数据为主的档案来讲,确保完整的语义来源,关键在于非结构化数据的结构化,其中对档案文本进行标注以及对档案内容、背景与结构进行著录是两种最主要的方法,所得的档案标注数据和档案著录数据是档案语义最重要的两类来源。

4.1.1　档案内容的转录与语义标注

对于很多传统纸质档案数字化转化而来的历史资源来讲,光学符号识别(Optical Character Recognition,OCR)是目前最为常用的文本转录方式。但 OCR 识别与转录之后的数据仅能支持机器对字符的识别与匹配,在数据含义上仍然做不到机器可理解,即这些转录后的数据仍然是非语义的。此时就需要对转录后的数据进行进一步的标注,从语义层面对其进行分析、序化、聚类和关联。与一般的档案资源著录(元数据)主要发生在文件层(document-level)不同,档案资源的语义标注强调下沉到实体层(item-level),也就是要对档案资源内容中的"事物"而非档案资源本身进行详细的描述。现有档案内容的标注常常采用手工建立标签(Tagging)的方式。例如,美国国家档案与文件署(National Archives and Records Administration, NARA)从 2011 年开始启动"公民档案员计划"(The Citizen Archivist Initiative)[41],鼓励公民通过添加标签、注释和翻译转录的方式,帮助实现 NARA 馆藏资源的结构化及其著录,NARA 为此还发布专门的标签政策[42]。自 2012 年以来,通过这种众包方式,公民贡献了数百万个标签、元数据、转录文本、视频字幕和数字图像等,为馆藏档案资源的内容理解与描述做出了重要贡献[43]。

除了众包方式的手工标注标签外,还有一些自动化的语义标注工具可供档案工作者使用,开展条目级的标注。例如由欧洲研究理事会资助的"交流与帝国:比较视角下的中华帝国"(Communication and Empire: Chinese Empires in Comparative Perspective)项目所开发的 MARKUS 自动化语义标注工具[44],目前支持中文和韩文两种语言文本中的语义实体自动标注,包括人物姓名、地名、时间、官衔、机构名称的标注等,同时还支持所有语言的自定义关键字列表或者标签的手工与批量标注。MARKUS 还与一系列概念模型或数据库建立了自动关联,如特定语言词典或特点领域的词汇表以及中国传记数据库(CBDB)、中国地理信息系统(CHGIS)等数据库,用于语言、领域知识、人名、地名等标注概念的参考。

无论是人工标注还是自动化标注,标注产出的众多数据都会成为档案资源描述数据的重要组成部分。这些标注数据有些可能与档案资源的著录项目重复或同义,如主题词与主题标签;有些可能与著录项目相关,如历史机构名称与档案生成机构等;有些可能是著录项目没有而对档案资源内容理解有重要补充,如历史人物名称、地名等。这些共同构成了描述档案资源的数据库,在经过进一步的语义关联后,会形成档案语义单元网络。

4.1.2　档案元数据的著录与语义增强

OCR 识别与语义标注一般用于档案内容语义的获取,对于档案背景语义与结构语义的获取则主要依靠档案元数据的获取与语义描述。元数据是最常用的档案描述和管理工具,也是目前信息管理系统中的档案结构化数据的最重要来源。元数据(metadata)是指数据的数据(data about data),最基本的功能就是定义和描述数据[30]12-13。元数据捕获数据的含义(meaning)部分,就是本文所述的语义网中的语义(semantic)一词。在纸质档案时期,元数据分散在案卷封皮、卷内目录等多个地方,需要重复记录。在电子文件或者信息管理系统时代,档案元数据具有了结构化、集中化、标准化等基本属性,一般以 XML 格式进行表达。例如,现有档案著录一般遵循的《编码档案著录规则》(Encoded Archival Description, EAD)就是一种基于 XML 模型的档案著录规则。但 XML 格式是一种隐式(Implicit)语义表达方式,也可视为不具有语义。因此,还要进一步对档案元数据进行语义增强,使元数据中蕴含的语义显性化。

例如,由欧盟发起的大型欧洲数字人文遗产资源项目"Europeana"存储了来自 3 700 多家欧洲图书馆、档案馆、博物馆和其他收藏机构超 5 800 万件的数字文化资源,这些资源都经过了基础元数据的著录,为了优化对大量元数据的检索和利用,Europeana 在语义层次上向元数据中添加新的信息,这一过程被称为"语义

增强”(Semantic Enrichment)[45] Europeana 的元数据语义增强过程主要分为三个阶段:①分析现有元数据集,选择参考数据集,制定元数据与参考数据集之间的匹配和关联规则;②将元数据项目及其取值与参考数据集中的字段和值进行匹配,并将参考数据集中的数据间关系自动添加到元数据集中;③将现有元数据集中没有而参考数据集中有的数据项及其取值,添加到元数据集中,包括语义相同或相似的概念、超类或子类概念等。经过这三个阶段的语义增强,档案元数据的语义得以显性化,能够更明确地建立起语义关联。

4.2 明确化的语义关联

捕获档案标注和著录数据,从结构化数据中获取语义,是档案语义组织的第一步。但任何语义都不可能独立存在,档案语义的内涵与边界需要更多的语义关系来界定,且语义之间的关系也决定了数据之间的关系。语义组织的核心就是建立语义关联,从而定义数据的关联,形成档案数据网络。语义关系本质上是概念之间的关系,概念之间的关系是由概念的外延决定的,反之这种关系又进一步影响了概念内涵的界定。因此,建立档案语义关系,就是找寻档案内容、结构和背景数据中所含概念之间的关系。本体是最适合完整表达档案资源中概念体系、严格定义并形式化表达概念与概念间关系的工具。本体(ontology),是语义网上用于描述资源元数据的数据字典(metadata vocabularies)。某个领域的本体就是关于该领域的一个公认的概念集,其中的概念含有公认的语义,这些语义通过概念之间的各种关联来体现[23]。

以芬兰国家级语义集体记忆平台 CultureSampo 为例,该平台希望通过语义关联实现对众多异质档案的整合,从而基于数字档案资源构建起一个完整的芬兰国家记忆。为此,CultureSampo 首先建立起一个国家级的数字资源本体 FinnONTO,将本国普遍使用的词表半自动化转化为轻量级本体,并通过不同领域专家之间的协作,在这些跨领域本体之间建立映射,最终形成一个全国性的大型本体——KOKO。KOKO 包含一个顶层本体 YSO(定义了 20 600 个概念)、一个博物馆领域本体 MAO(定义了 6 800 个概念)、一个农业林业领域本体 AFO(定义了 5 500 个概念)、一个应用艺术领域本体 TAO(定义了 2 600 个概念)和一个摄影本体论 VALO(定义了 1 900 个概念)[46]。这些本体为 CultureSampo 对各类数字资源的标注数据和元数据之间建立关联提供了基础框架和依据。

4.3 形式化的语义表达

语义的表达方式按照机器能否直接理解分为隐式表达、非形式化表达和形式化表达三种[24]。一般档案文本内容都采用自然语言进行表达,是一种非形式化的语义表达方式,人可以阅读并理解,但机器无法理解其中的语义和语义关系,即不具有语义;一般档案元数据都采用 XML 语义进行表达,尚处于语法层次,描述的是数据的结构而非数据的含义,因此是一种隐式语义表达方法,也可以说不具有语义。因此,要想使档案语义为机器可理解,就要用形式化的语言重新表达档案元数据。形式化表达的语义是一种模型论语义(Model Theoretic Semantics),即用一定结构和模型的“声明”来定义语义。RDF 的三元组(一个三元组就是一个声明)就是一种模型论语义表达即形式化表达方法。档案语义从隐式语义表达方式向形式化语义表达方式的转化,就是档案资源标注和著录数据及其语义描述数据从 XML 文档向 RDF 文档的转化。RDF 可以基于 XML 语法,这就为现有很多以 XML 格式存储的档案元数据转化为 RDF 格式提供了可能。

英国 Archives Hub 的关联数据项目就是通过档案元数据的 RDF 转化,实现了档案语义的形式化表达。Archives Hub 是为社会各层级用户提供对英国境内 363 家机构的档案著录数据的交叉检索的非盈利机构,其本身并不保管任何档案资源,但存储了大约 174 万余个馆藏档案汇集的元数据,所有的这些元数据都面向社会用户提供检索[47]。为了使机器自动理解档案数据的语义并智能化服务用户的检索需求,Archives Hub 专门启动了 Locah(Linked Open Copac Archives Hub)项目,探索出档案语义形式化表达的基本步骤:①构建档案关联数据模型(本体);②检索和复用已有的词表(或本体),填充档案关联数据模型中缺少的概念;③为档案元数据添加 URI;④将档案元数据从 EAD 数据转化成 RDF XSLT 样式表;⑤发布档案关联元数据; 创建关联数据视图; 使用 SPARQL 语言进行数据的语义关联等[48]。其中,将档案元数据从 EAD 格式转化成 RDF XSLT 样式表是显性描述元数据语义的关键一步。RDF XSLT 样式表能够对①中构建的数据模型进行封装,从而提供一种简单的、标

准化的、可重用的、档案元数据形式化转换为关联数据的方案[49]。

综上所述，通过语义标注和元数据语义增强，档案的内容、背景与结构得以转化为隐藏语义的结构化数据，通过明确化的概念模型建立起这些数据之间的语义关系，再通过机器可理解的形式化语言来表达这些语义和语义关系，从而构建起一个富含结构、语义和关联的机器可理解的数据网络，这是档案语义组织的基本原理。而要从基本原理到不同类型档案语义组织的实现，还要借助更多语义技术与工具，基于档案资源实际状态和档案业务场景继续深入探索。

5 档案语义组织的基本原则

档案语义组织在不同的档案资源状态和档案业务场景支持下，可能会有不同的实践路径，但都应遵循档案语义组织的基本原则，这些原则也体现了档案语义组织与其他信息资源语义组织的区别于联系。一方面，资源特性的不同导致的语义来源数据的不同，因此档案的语义需要完整来源于各层级的内容、背景与结构；另一方面，档案语义组织在方法上继承了信息资源语义组织的共性方法，但共性方法应用在特性资源上，就需要同时尊重档案资源特性与语义组织的基本规律。

5.1 语义完整原则

档案语义组织的第一步是获取语义，在语义获取过程中应遵循语义完整原则，包括：

（1）档案的全宗－类别－案卷－文件等各级别的内容、背景和结构都应被完整著录，上下级别之间在内容、背景和结构上的继承关系应被充分考虑。

（2）各级别的著录项目与格式应遵循一定标准设定，以实现档案的内容、背景和结构数据的充分结构化，并对结构化数据进行充分的语义描述，包括所有数据的语义内容、结构、格式和关系等，以支持数据含义的注解与抽象化定义。

（3）在分析档案的语义关联时，无论是某档案向内语义组织向外语义组织，都应充分考虑内容、背景、结构各自内部语义的关系，以及内容、背景、结构之间语义的关系。

（4）根据不同服务对象与目的，档案语义组织框架的中心可以是内容语义，也可以是背景语义，结构语义一般围绕内容语义或背景语义进行关联。例如，当面向历史或人文研究时，档案语义组织的中心应选择内容语义，背景和机构语义可以为内容语义的理解与相关关系的建立提供关联支持；当面向业务支持时，档案语义组织的中心应选择背景语义，内容语义和结构语义可为机构背景、业务背景、程序背景与文件背景等提供支持。

5.2 链式关联原则

链式关联原则是指在建立档案内容、背景与结构语义之间的关联时，应尊重且遵循档案内容、背景和结构中的链条式逻辑。链式关联原则中的“链”包括：

（1）内容逻辑链。对档案内容语义的关联主要依据内容逻辑链，包括档案内容中所涉时间、地点、人物、事件（主题）、实物要素的各自变化与要素之间的关系及关系的变化，常见于时间轴、位置变迁、人物网络、事件叙事、实物变化等单层逻辑，以及基于时间轴的位置变迁、基于时间轴或位置变迁的人物网络、基于事件叙事的人物网络或实物变化等双层或多层逻辑。这些内容逻辑往往符合历史或人文研究的科学逻辑。基于内容逻辑链的档案语义组织以内容语义为中心。

（2）机构职能链。对不同机构不同档案汇集之间的向外语义组织，可以依据机构之间的职能关系分析档案汇集之间的关系。因此，基于机构职能链的语义关联主要发生在档案汇集层，以背景语义尤其是机构背景语义为中心，对档案内容、背景与结构语义之间可能存在的关系在不同档案汇集之间进行关联。

（3）业务流程链。对于同一业务流程中所产生的文件之间的语义关联，可以依据业务流程链，即文件生成时所参与的业务活动在业务流程中所处环节和位置，决定了文件之间的关系。在日趋复杂的业务环境中，一个业务流程链可能在一个部门完成，也可能跨越一个机构的多个部门，甚至跨越多个机构。因此根据业务流程链开展的语义关联，可能发生在同一文件系列中的不同文件之间，可能发生在同一档案汇集内的不同文件系列之间，也可能发生在不同档案汇集的某些文件之间。基于业务流程链的档案资源语义关联将以背景语义尤其是业务背景语义作为中心，寻找内容、背景、结构中更多的相关关系。

（4）文件生命周期链。文件生命周期赋予了文件动态性，表现为包括文件生成、转递、归档、长期保存和

利用的程序链。这种程序链从管理角度说明了文件在机构职能、业务流程中所处的位置，而电子信息系统能够为这些程序留下数据痕迹。文件生命周期链对于一份文件的内容、背景与结构语义之间的关系的建立至关重要，尤其是文件、责任者、系统环境等之间的关系。基于文件生命周期逻辑的档案资源语义组织可以程序背景语义为中心，寻找结构语义或内容语义与其的关系。

(5)以上四种链式逻辑的组合。对于档案语义组织来讲，向内组织、向外组织同样重要，将内容逻辑链、机构职能链、业务流程链、文件生命周期链集合起来开展语义组织，能够发现更多的语义中心和关系网络，实现档案汇集之间-档案汇集内文件系列之间-文件系列内文件之间-文件内数据之间的多层网络化关联。

5.3 网络多维原则

网络多维原则是指经档案语义组织所形成的档案数据网络应该是一个非唯一中心的多维网络，能够支持多角度、多维度的检索查询与智能化服务。在继承和发展传统的强调等级关系的层累制组织方法基础上，档案语义组织更强调相关关系的网络化组织。在这个网络中，没有唯一的中心，而是不同层级中的内容、背景和结构中的任何一个语义单元都可以成为中心，而按照链式关联的原则向外发散与其他语义单元建立关联。网络多维原则中的“多维”包括两种基本维度和一种交叉维度：

(1)基于同一层级档案内容、背景、结构语义之间关联的横向数据网络。

(2)基于上下层级之间档案内容、背景、结构语义之间关联的纵向数据网络。

(3)基于不同层级之间档案内容、背景、结构语义之间关联的交叉数据网络。

多种维度的数据网络的建立能够为用户提供不同的检索点，为档案在同一层级、上下层级、间隔层级之间的扩检、缩检、改检等提供支持，并为档案资源可视化范围与结构的灵活多变提供了可能。多维、去中心的网络化组织是档案语义组织的基本理念，也是最终档案资源语义组织所构建出的富含语义的数据网络的基本特征。

6 结语

为实现档案数据化“机器可理解”“机器可操作”的核心目标，档案语义组织从数据的含义以及数据含义之间的关联出发，对档案内容、背景与结构数据进行语义层面的序化、聚类与关联，形成基于多级著录成果的多级语义网络，将档案汇集内各文件之间和文件的内容、背景与结构之间关联起来，也将不同档案汇集以及档案汇集与其他领域数据集之间关联起来，构建起富含语义与语义关系的领域数据网络，以支持未来更多元技术背景下的档案细颗粒开发与智能化应用。

档案语义组织不是一个从零开始的全新过程，既有的档案信息组织工具、方法与成果，能够为档案语义组织奠定良好基础。例如，已有的档案著录规则经过本体化后，能够支持语义著录；已有的档案著录和标注数据，经过形式化表达后，能够支持档案关联数据集的建立等。这些都是档案语义组织在实践层面的具化路径。但总体上讲，档案语义组织仍然要经过语义含义明确化、编码形式化、关联链接化的过程。未来可能出现的更多语义技术将为档案语义组织带来新的方法，但档案语义组织的原则应当始终遵守，只有在尊重档案资源特性和管理专业性的基础上，才能探索出更多的档案语义组织落地方案。

参考文献：

[1] 钱毅. 技术变迁环境下档案对象管理空间演化初探[J]. 档案学通讯,2018(2):10-14.

[2] 赵跃. 大数据时代档案数据化的前景展望：意义与困境[J]. 档案学研究,2019(5):52-60.

[3] Negroponte Nicholas. Being digital [M]. New York: Vintage Books,1996.

[4] 姜浩. 数据化 由内而外的智能[M]. 北京：中国传媒大学,2017.

[5] 陶俊. 词表语义组织研究的演进(1998-2018)[J]. 图书情报工作,2018(21):140-148.

[6] 谈国新，侯西龙，庄文杰. 非物质文化遗产多媒体资源语义组织研究[J]. 图书馆学研究, 2017(24):44-54.

[7] 王曰芬，邢梦婷. 面向政府决策需求的社会舆情信息语义组织研究[J]. 现代图书情报技术, 2016, 32(7):21-31.

[8] 马雨萌，郭进京，王昉. e-Science环境下科学数据语义组织模型框架研究[J]. 现代图书情报技术, 2015(7):48-57.

[9] 袁艳. 科技报告中的知识发现研究[J]. 图书馆界,2017(5):82-84.

[10] 丁恒，陆伟. 标准文献知识服务系统设计与实现[J]. 数据分析与知识发现, 2016, 32(7-8):120-128.

[11] 毕传龙. 大数据时代民俗文化资源的数字化[J]. 民族艺术研究,2016(3):87-93.

[12] 马寅源. 关联数据应用于档案知识服务的 SWOT 分析及策略[J]. 档案与建设,2017(2):17 - 20.
[13] 裘丽. 后保管时代下构建档案知识服务模式探索[J]. 云南档案,2015(9):52 - 55.
[14] 林周佳. 档案的语义级检索技术研究[J]. 档案与建设,2007(9):26 - 27.
[15] 任妍,庞宇飞,荆欣. 全媒体档案信息资源语义组织与服务研究[J]. 档案管理,2019, 237(2):37 - 38.
[16] 沃尔钦科夫. 苏联档案工作的组织(在 1956 年 12 月 22 日全国档案工作会议上的报告)[J]. 档案工作,1957(2):5 - 9.
[17] 和宝荣, 陈兆祦, 松世勤. 文书档案工作基本知识讲座(提纲)——第四章 档案的整理[J]. 档案工作, 1980(4):27 - 33.
[18] 周铭. 殊途同归:档案分类法与主题法研究[J]. 四川档案, 2000(1):12 - 14.
[19] 邓绍兴.《中国档案分类法》是一部具有我国特色的档案分类法[J]. 北京档案, 1996(9):20 - 23.
[20] 曾娜. 网络档案信息资源组织研究[J]. 档案学通讯, 2010(1):45 - 49.
[21] 赵屹. 网络档案信息资源组织方式[J]. 科技文献信息管理, 2003(4):15 - 19.
[22] 李海军. 档案信息转化为"档案知识"的技术框架探讨[J]. 山西档案, 2007(1):28 - 30.
[23] 王兰成. 论知识集成环境下的档案信息组织与检索发展[J]. 档案学研究, 2008(5):45 - 50.
[24] 吕元智. 数字档案资源知识"关联"组织研究[J]. 档案学研究, 2012(6):46 - 50.
[25] 吕元智. 数字档案资源体系的语义互操作实现研究[J]. 档案学通讯, 2013(5):53 - 57.
[26] 吕元智. 数字档案资源跨媒体语义检索实现框架与关键问题研究[J]. 档案学研究, 2014(2):65 - 70.
[27] 吕元智. 数字档案资源跨媒体语义关联聚合实现策略研究[J]. 档案学研究, 2015(5):60 - 65.
[28] 第十三届国际档案大会文件报告集[C]. 北京:中国档案出版社,1997.
[29] 梁孟华. 基于开放关联数据的数字档案资源跨媒体知识链接研究[J]. 档案学研究,2015(4):111 - 116.
[30] 冯惠玲. 电子文件管理 100 问[M]. 北京:中国人民大学,2014.
[31] 戎军涛. 学术文献内容知识元语义描述模型研究[J]. 情报科学,2019(7):30 - 35.
[32] 杜協昌, 項潔. 臺灣歷史數位圖書館[EB/OL]. [2021 - 02 - 17]. http://doi.org/10.6681/NTURCDH.DB_THDL/Text.
[33] 贺德方, 曾建勋. 基于语义的馆藏资源深度聚合研究[J]. 中国图书馆学报,2012, 38(4):79 - 87.
[34] DURANTI L. The archival bond [J]. Archives and museum informatics, 1997, 11: 213 - 218.
[35] 杜協昌, 項潔. 臺灣歷史數位圖書館[EB/OL]. [2021 - 02 - 17]. http://thdl.ntu.edu.tw/THDL/RetrieveSVG.php? filename = ntu-2252926-0080500806-0000840.txt.
[36] OLIVER W H. Archival arrangement - Five different operations at five different levels [J]. The American archivist, 1964, 27(1):21 - 42.
[37] 张正强, 卞刚. 现代档案著录的原则与原理[J]. 中国档案,1999(10):39 - 41.
[38] ISAD(G): General international standard archival description [S]. Second edition. ICA,1999:36.
[39] 马寅源. 国内外档案多级著录的比较研究[J]. 档案学研究, 2017(02):53 - 58.
[40] DAMA 国际. DAMA 数据管理知识体系指南[M]. DAMA 中国分会翻译组,译. 北京:机械工业出版社,2020.
[41] National Archives. Citizen archivist dashboard [EB/OL]. [2021 - 02 - 27]. https://www.archives.gov/citizen-archivist.
[42] Citizen archivist dashboard. Citizen contribution policy [EB/OL]. [2021 - 01 - 27]. http://www.archives.gov/citizen-archivists/resources/tagging-policy.
[43] ANDREW W. Citizen archivist dashboard/Improving access to historical records through crowdsourcing [EB/OL]. [2021 - 01 - 27]. https://www.citizenscience.gov/citizen-archivist/#.
[44] MARKUS [EB/OL]. [2021 - 01 - 27]. https://dh.chinese-empires.eu/markus/beta/.
[45] HUGO M. Europeana semantic enrichment framework [EB/OL]. [2021 - 02 - 01]. https://docs.google.com/document/d/1JvjrWMTpMIH7WnuieNqcT0zpJAXUPo6x4uMBj1pEx0Y/edit.
[46] HYVÖNEN E, VILJANEN K, TUOMINEN J, et al. Building a national semantic Web ontology and ontology service infrastructure-the FinnONTO approach [A]//The semantic Web: research and applications. Berlin: Springer, 2008: 95 - 109.
[47] WIKIPEDIA. Archives hub [EB/OL]. [2021 - 01 - 28]. https://en.wikipedia.org/ wiki/ Archives_Hub.
[48] Linking Lives. About locah [EB/OL]. [2021 - 01 - 28]. http://linkinglives.archiveshub.ac.uk/about-locah/.
[49] ADRIAN S. Final froduct post: Archives Hub EAD to RDF XSLT stylesheet [EB/OL]. [2021 - 01 - 28]. http://locah.archiveshub.ac.uk/tag/linkeddata/.

作者贡献说明:

祁天娇:负责论文撰写;
冯惠玲:论文指导。

祁天娇,冯惠玲. 档案数据化过程中语义组织的内涵、特点与原理解析[J]. 图书情报工作,2021,65(9):3-15.

The Connotation, Characteristics and Principle Analysis of Semantic Organization in the Process of Archival Datalization

Qi Tianjiao　Feng Huiling
School of Information Resource Management, Renmin University of China, Beijing 100872

Abstract: [**Purpose/significance**] In the stage of archival datalization, archival utilization and service need to meet the new needs on the data level, breaking through the limitation of page level reading and file level using. This requests a new semantic organization mode for archives, supporting deep mining and analysis on the data in archival content, context and structure, to prepare resource, methods and technologies for archival value enrichment, resource development and intelligent knowledge services. [**Method/process**] Based on the phase characteristics of archival datalization, through literature investigation and cases study, this paper analyzed the basic connotation of archival semantic, semantic relation and semantic organization, compared the differences and features of archival semantic organization with the semantic organization of other information resources, and explored the theoretical framework of archival inward and outward semantic organization under the principles of semantic integrity, chain association and multi-dimensional network. [**Result/conclusion**] Archival semantic organization is carries out based on the meaning and linkage of data, aimed at finding the semantic relation from the content, background and structure data of archives. The archival semantic organization is the key link to realize the archival datalization and the key step to realize the archival machine-understandable and machine-operable. Through archival semantic organization, the originally scattered, disturbed and field-dependent archival content, background and structure data, could have clear definition, formal expression and associated links. Archival data could be machine-understandable and machine-operable. It is possible for archival resources to be organized, preserved and used automatically with linkages, thus eventually support the intelligent acquisition and utilization of archives based on human-machine and machine-machine interaction.

Keywords: archives　archival datalization　semantics　semantic relation　semantic organization